Credit card default analysis: Final report

Problem Statement

This Capstone Project centers on the critical challenge of predicting credit card default risk, employing a dataset from April to September 2005. It meticulously explores the multifaceted relationship between demographic attributes, credit data, and payment histories to construct a predictive model. Drawing inspiration from seminal works in predictive analytics and machine learning, such as Altman's discriminant analysis and Breiman's ensemble learning techniques, the project emphasizes the importance of advanced statistical methods in financial risk assessment. The comprehensive analysis extends beyond model accuracy, delving into demographic analyses and feature engineering to unearth deeper insights into credit behavior and risk factors. Leveraging a rich dataset from the UCI Machine Learning Repository and guided by literature on credit scoring, the project adopts a methodical approach encompassing data cleaning, exploratory data analysis, and rigorous model evaluation. The culmination of this endeavor is a set of deliverables including a well-documented codebase, this report detailing methodologies and findings. This project not only aims to enhance decision-making for credit card companies but also sets the stage for future exploration in model optimization, data enrichment, and interactive data visualization, ensuring its relevance and adaptability in the evolving landscape of machine learning and financial analytics.

Dataset and data wrangling

In the "Data Wrangling" step for this project, the initial step involves importing essential libraries for data manipulation and visualization, followed by reading the dataset into a pandas DataFrame from an Excel file. This foundational step sets the stage for data cleaning and exploration. Subsequently, unnecessary columns, such as 'ID', are removed to streamline the dataset, enhancing its relevance and manageability for analysis. The renaming of columns, particularly those related to repayment statuses, bill amounts, and previous payments, is carried out to improve the dataset's readability and facilitate easier data handling. Furthermore, the notebook includes a detailed examination of outliers using boxplots for both bill amounts and previous payments. This critical step aids in identifying data points that significantly deviate from the norm, allowing for informed decisions on whether to adjust or remove these outliers to ensure the dataset's integrity and reliability for further analysis and modeling.

List of variables

- Repayment Status (April to September): Represented as `repayment_status_april` to `repayment_status_sept`, these variables likely indicate the repayment status of the credit card holder for each month. A specific coding system (e.g., -1 for pay duly, 1 for payment delay for one month, etc.)
- Bill Amount (April to September): `bill_april` to `bill_sept` variables represent the bill statement amount for each month.
- Previous Payment (April to September): `previous_payment_april` to `previous_payment_sept` denote the amount of previous payment for each month.
- Limit Balance: The `LIMIT_BAL` variable represents the amount of given credit.
- Sex: A categorical variable indicating the gender of the credit card holder.
- Education: The `EDUCATION` variable, with values being replaced and categorized, likely represents the education level of the credit card holder.
- Default Payment Next Month: The `default` variable indicates whether the credit card holder defaulted the following month. This is the target variable for predictive modeling.

EDA

Demographic variables

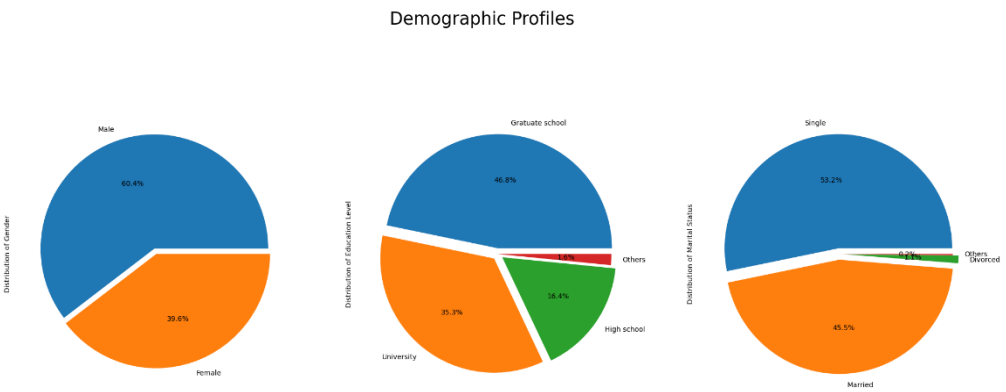
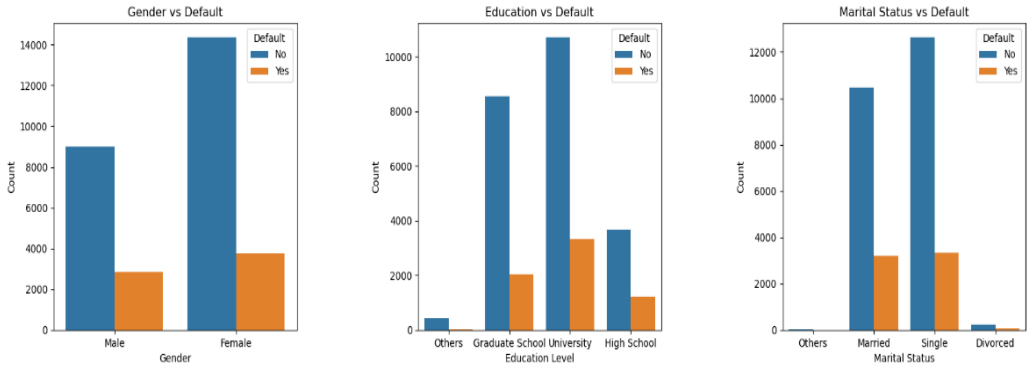


Figure 1:Demographic variables vs Default

The pie charts collectively offer a comprehensive view of the demographic makeup of the dataset, highlighting gender, educational background, and marital status distributions. The gender distribution chart indicates a specific balance between male and female clients, suggesting

potential directions for gender-targeted financial services. The education chart reveals a predominant representation of clients with higher education levels such as 'Graduate School' and 'University', pointing towards a clientele that may possess greater financial literacy and distinct credit needs. Lastly, the marital status distribution, with segments for 'Married', 'Single', 'Divorced', and 'Others', underscores the diverse familial and social contexts within which clients operate, impacting their financial behaviors and credit risk profiles. These insights are crucial for tailoring credit policies, products, and marketing strategies to better meet the varied needs of the client base.

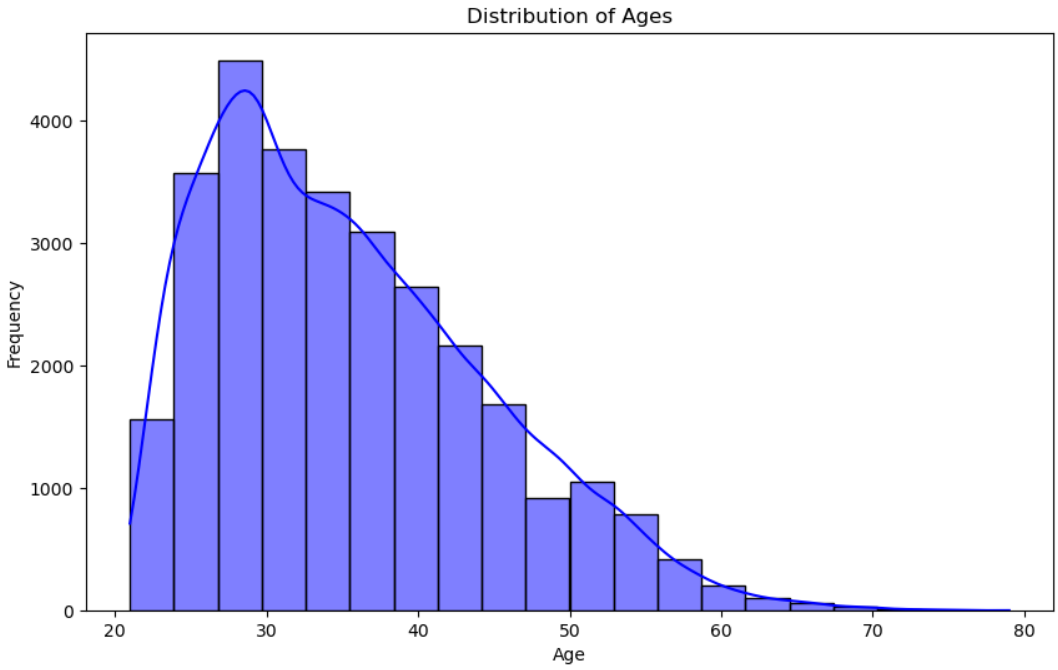
Figure 2: Demographic variables vs Default



The intended bar charts, showcasing the relationships between demographic factors (Gender, Education Level, and Marital Status) and default rates, are critical in understanding the dynamics of financial behavior across different groups. For Gender, the analysis might reveal distinct patterns in default rates, possibly indicating varying financial management styles or obligations between males and females. The Education chart could provide insights into how educational backgrounds correlate with financial stability, with the potential to highlight higher default rates among certain educational levels. Lastly, the Marital Status chart would have offered a glimpse into how life stages and social constructs around marital status impact financial decision-making and risk. These insights are invaluable for financial institutions to develop nuanced risk models and to tailor financial advice, product offerings, and marketing strategies to cater to the diverse needs of their clientele.

Age

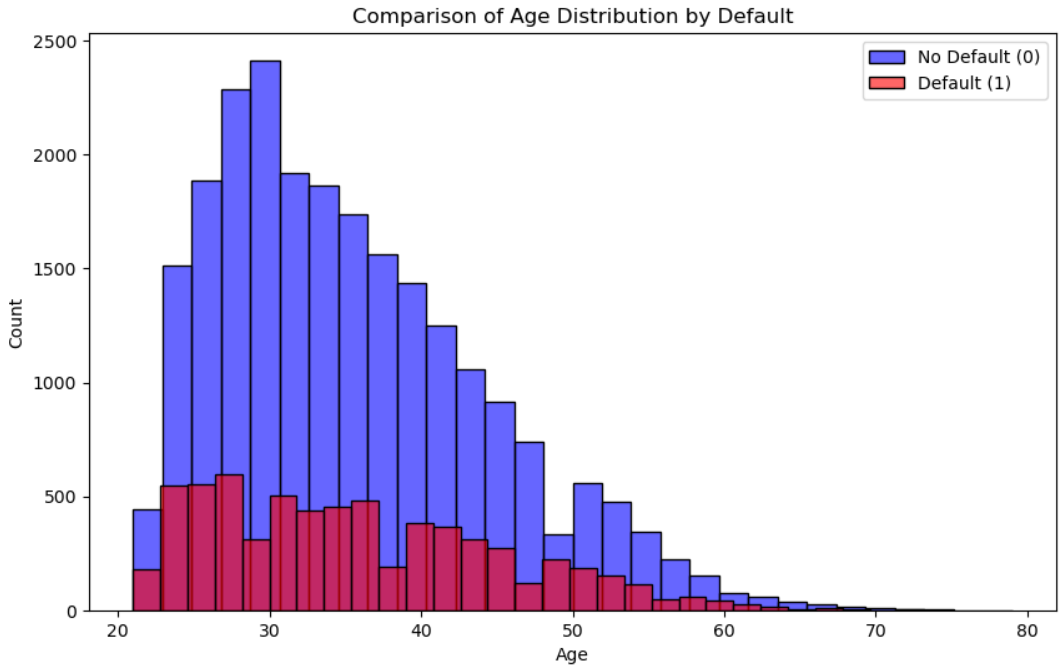
Figure 3: Age distribution



The age distribution histogram, augmented by a Kernel Density Estimate (KDE), offers a comprehensive view of the client base's demographic structure, pinpointing the predominant age groups within the dataset. Typically characterized by its skewness, the distribution may lean towards younger individuals in the case of a right skew, indicative of a client base that is potentially more receptive to new financial products and services. Alternatively, a left-skewed distribution suggests a predominance of older clients, who may have different financial needs and risk considerations.

The KDE line, by smoothing the age distribution, aids in identifying subtle patterns such as bimodal distributions, which could imply the existence of distinct demographic subgroups with varying financial behaviors. This demographic insight is pivotal for financial institutions as it underpins the development of tailored financial solutions. By aligning products, services, and marketing strategies with the age-related preferences and risk profiles of their clientele, institutions can enhance client engagement and satisfaction.

Figure 4: Age distribution by default



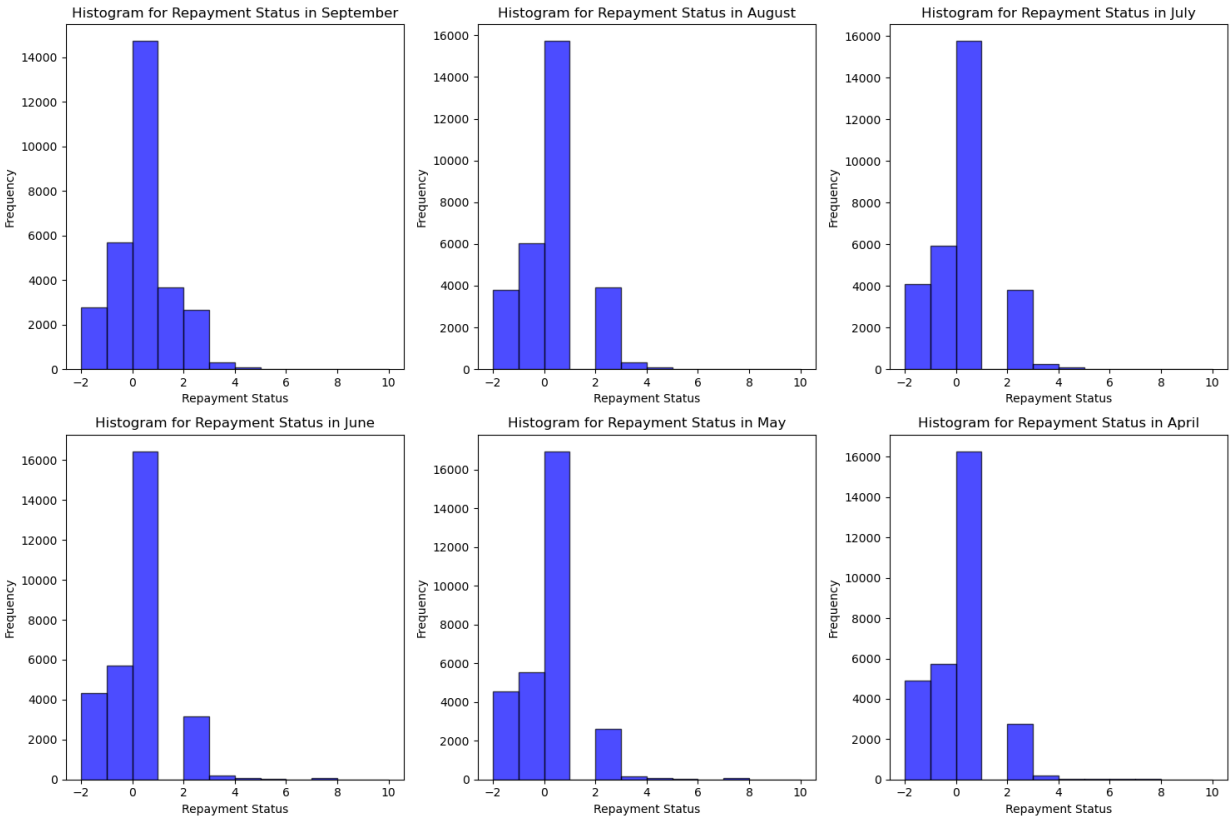
The histogram graph effectively visualizes the age distributions of clients based on their default status on credit payments, offering insightful contrasts between those who have defaulted (Default 1) and those who have not (No Default 0). A key observation is the similarity in the shape of the age distributions for both groups, with a pronounced concentration of clients in the younger age brackets, particularly between 20 and 40 years. This age range, peaking around the mid-20s to early 30s, signifies a higher engagement with credit facilities among younger individuals within the dataset. The higher counts in the 'No Default (0)' category across most age bins indicate that a larger proportion of clients in various age groups have maintained a good track record of payments. The overlay of the 'No Default (0)' distribution in blue over the 'Default (1)' distribution in red highlights significant overlaps, especially among the younger demographics, suggesting the presence of both defaulting and non-defaulting individuals within these age groups.

The graphical representation also points to a decline in counts as age increases, suggesting either a reduced number of older individuals in the dataset or potentially a decrease in credit usage or default risk among older age groups. For financial institutions, such insights are valuable for risk management, potentially guiding strategies for credit policy adjustments or targeted educational initiatives aimed at younger clients to mitigate default risks.

Overall, this visualization underscores the significance of age in credit risk analysis while also reminding us that default risk is influenced by a multitude of factors and cannot be solely attributed to age.

Repayment status

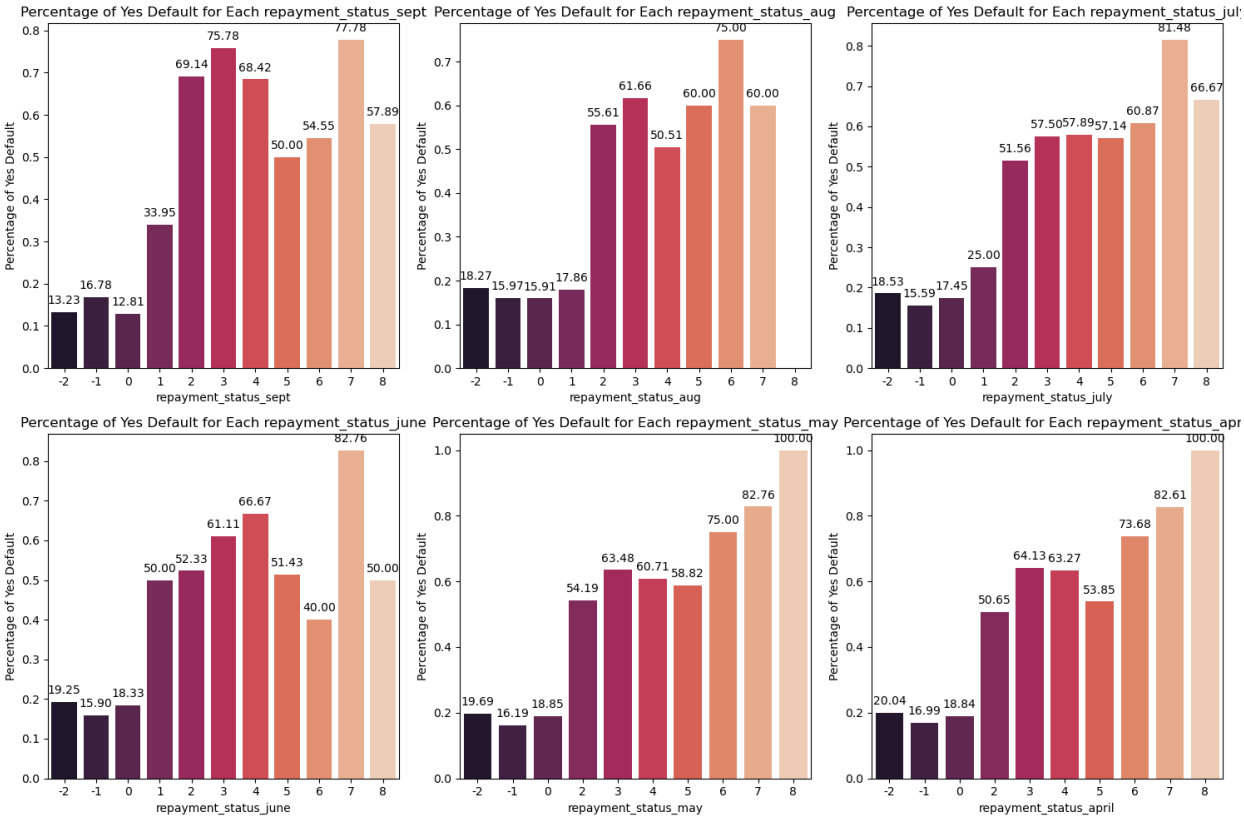
Figure 5: Repayment status distribution



The series of histograms offers a comprehensive visual examination of the repayment statuses across six consecutive months, from September to April, within the dataset. Each histogram corresponds to a month, arranged in a way that allows for an easy comparison of repayment behaviors over time.

These histograms showcasing the repayment statuses across several months reveal significant insights into the payment behaviors of individuals within the dataset. A notable observation is the diversity in repayment statuses, suggesting a broad spectrum of financial behavior among credit users. Certain statuses are more prevalent, as indicated by the taller bars in the histograms, pointing towards common trends such as a general adherence to payment schedules or minor delays in payments.

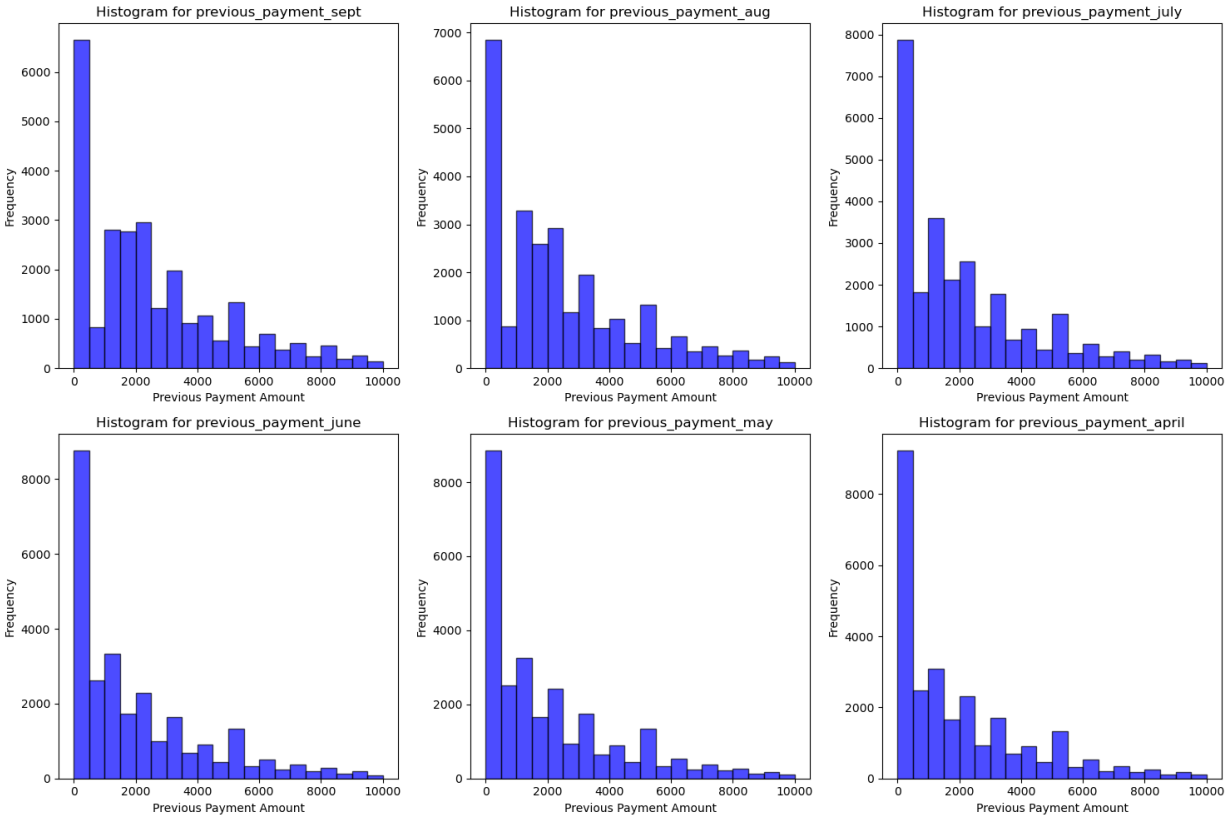
Figure 6:Repayment status distribution by default



The bar graphs that detail the percentage of defaults across various repayment statuses for each month, spanning from September to April, elucidate key insights into the dynamics of financial behavior and default risk. These visualizations reveal a notable variability in default rates among different repayment statuses, underscoring the pivotal role of payment behavior in determining the likelihood of default. Particularly striking are certain repayment categories that exhibit markedly higher default percentages, signaling them as high-risk zones warranting closer scrutiny and targeted intervention by financial institutions. This nuanced understanding of repayment behavior's impact on default risk not only reaffirms the critical value of repayment history in credit risk assessments but also offers strategic insights.

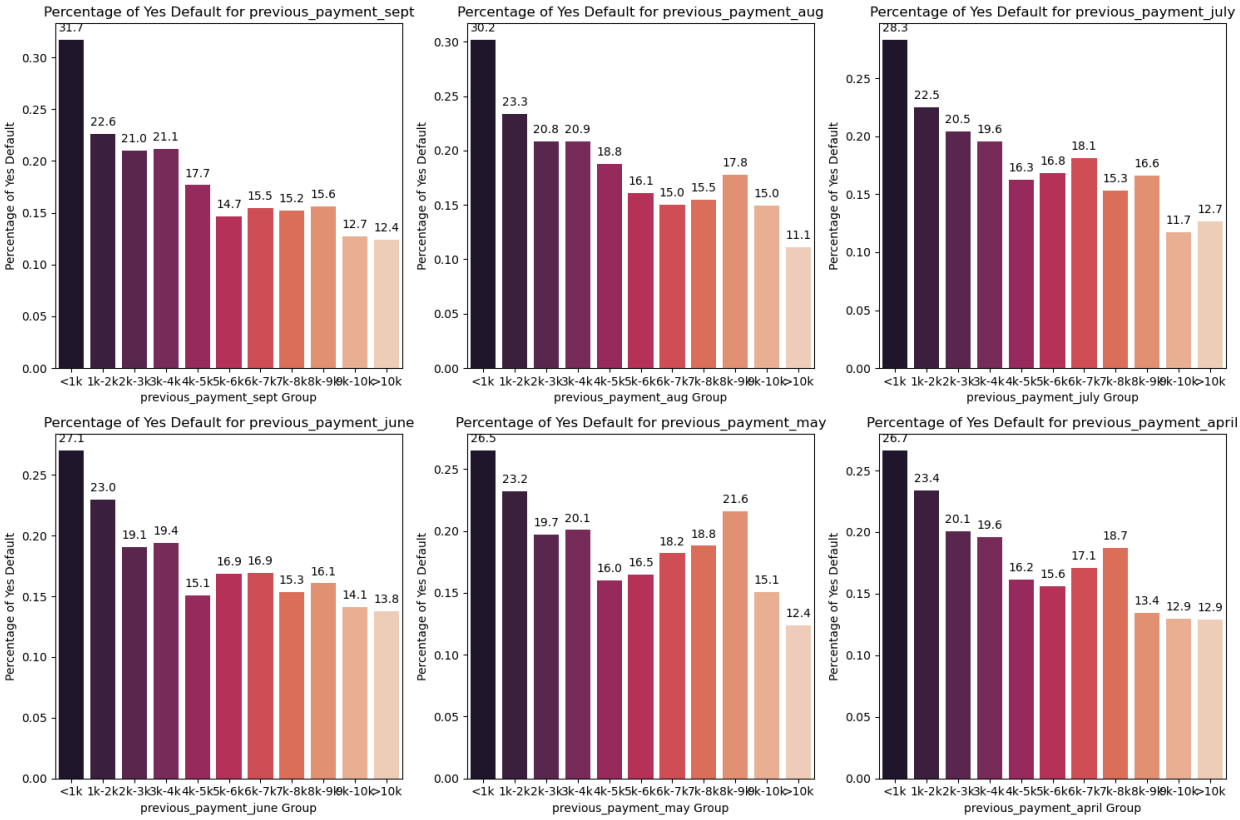
Previous payment

Figure 7: Previous payment distribution



The histograms generated for the 'previous_payment' variables from September to April reveal a consistent pattern in payment behavior among account holders, characterized by a right-skewed distribution. This suggests that most account holders tend to make smaller payments, with the majority falling within the 0 to 500 range, indicative of minimum or nominal payment amounts. Despite this trend towards smaller payments, there is a noticeable presence of payments extending up to and beyond 10,000, highlighting a segment of account holders who engage in higher payment transactions. The consistency in the shape and distribution of these histograms across the different months indicates a stable payment behavior over time. However, the presence of significant outliers or larger payments suggests variability in the account holders' financial behavior or circumstances, warranting further investigation to understand the underlying factors driving these larger transactions.

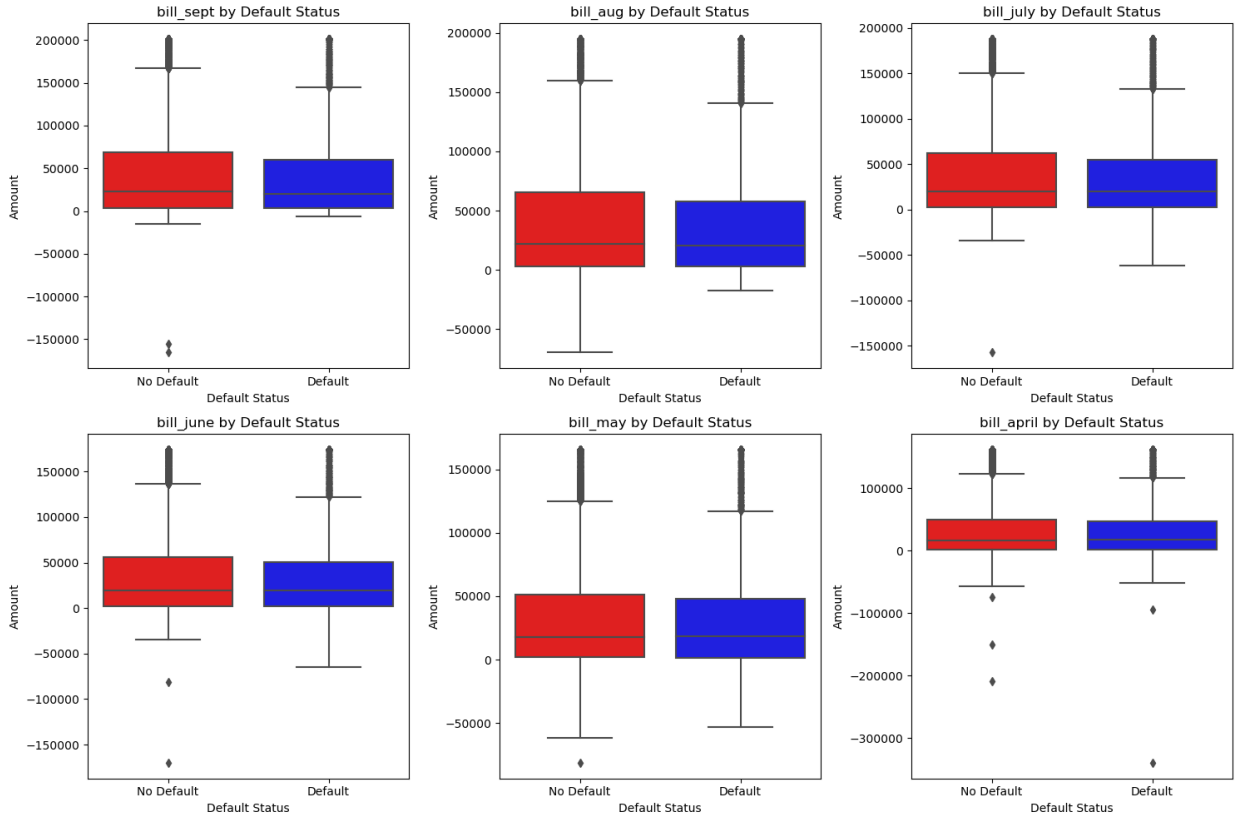
Figure 8: Previous payment distribution by default



This bar plots generated to examine the relationship between previous payment amounts and default rates reveal nuanced insights into payment behavior and financial risk. The variation in default percentages across different payment amount groups suggests a potential correlation between how much individuals pay in previous months and their likelihood of defaulting. Notably, lower payment groups tend to show more consistent default rates, whereas higher payment categories exhibit greater fluctuations. These fluctuations could be attributed to the lesser number of accounts in these higher payment brackets, which makes the default rate more sensitive to individual defaults. Furthermore, the annotations on each bar provide clear, quantitative insights into the default rates within each group, enhancing the interpretability of the data. Despite the variability, the persistence of these patterns across different months suggests a stable relationship between past payment behaviors and default risk, indicating that payment history could be a significant predictor of financial distress. This analysis underscores the complexity of credit risk assessment and the importance of considering a range of behavioral factors when evaluating default risk.

Bill amt

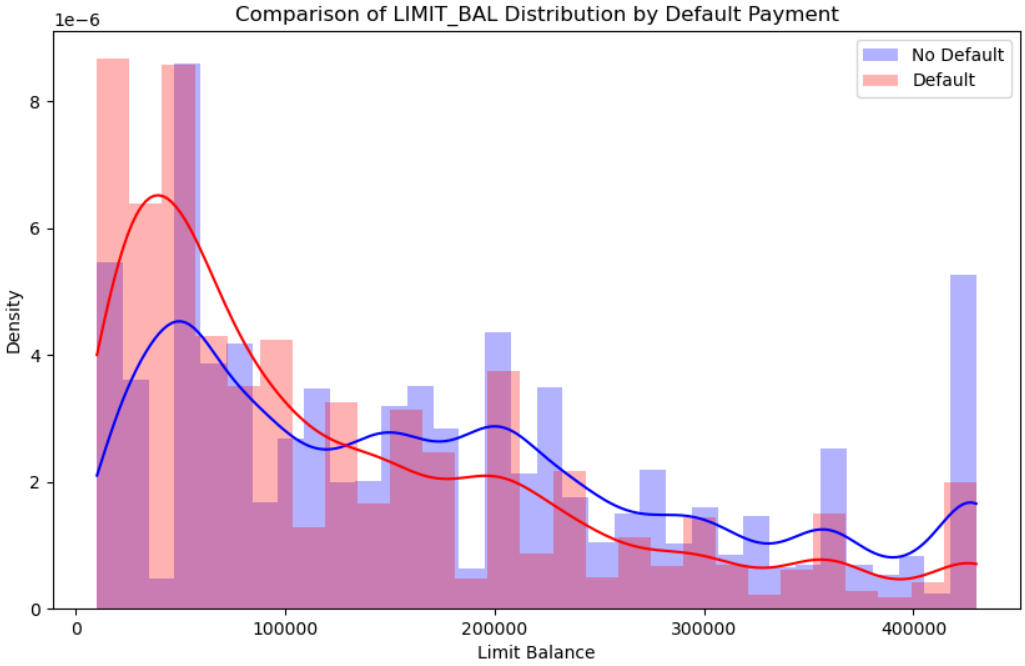
Figure 9: Bill amount by default



The box plots vividly illustrate the distribution of bill amounts across six months, revealing a consistent pattern of wide variability and numerous outliers in both default and non-default groups. Despite the presence of outliers indicating significantly high bill amounts in some accounts, the median values for default and non-default groups are relatively similar across all months. This similarity suggests that the mere level of bill amounts may not be a definitive indicator of default risk. The plots, differentiated by red for non-defaults and blue for defaults, underscore the complexity of financial behaviors and the need for a multifaceted approach in assessing credit risk. The right-skewed distribution observed across all variables further emphasizes the predominance of lower bill amounts among the majority of accounts, with a smaller fraction incurring higher charges.

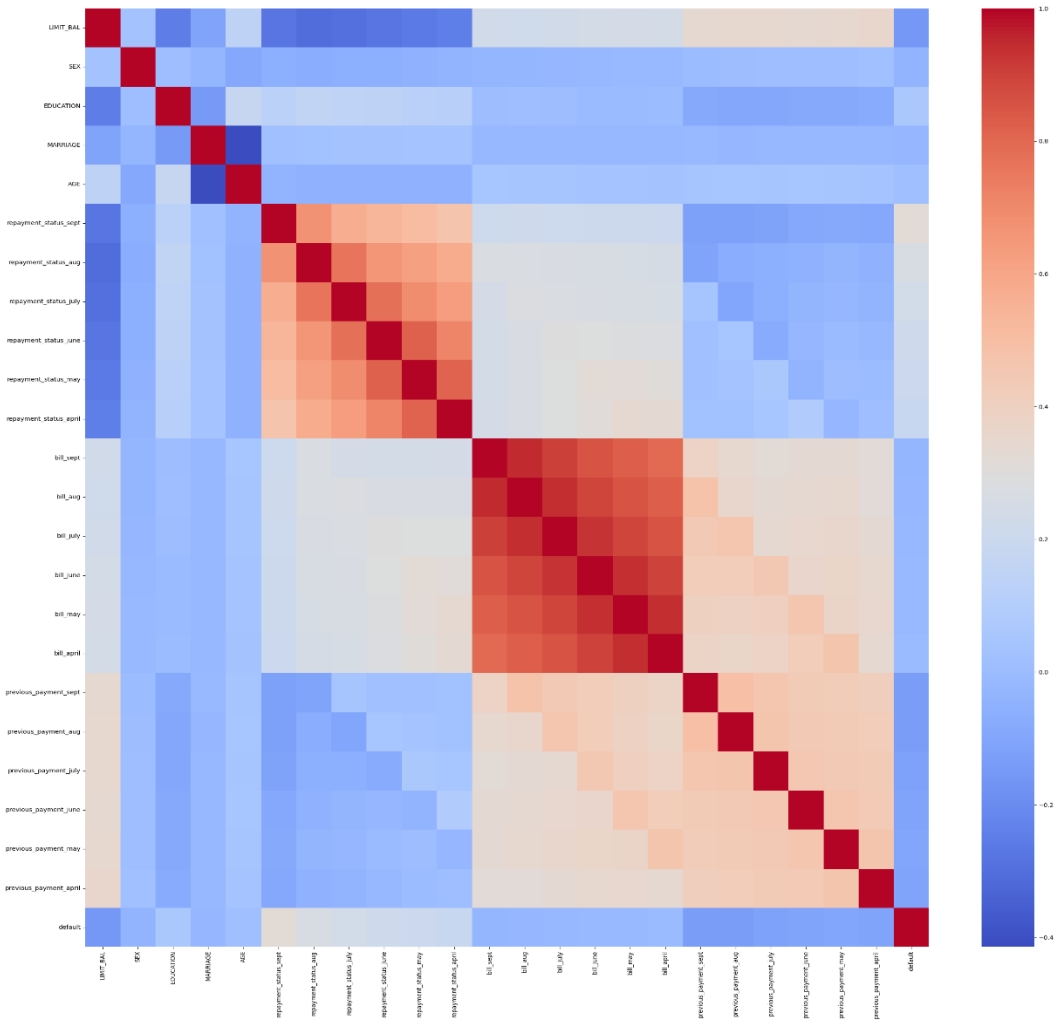
Limit ball

Figure 10: Limit ball distribution by default



The histograms and overlaid density plots, with default accounts depicted in red and non-default accounts in blue, highlight a clear distinction in the distribution of credit limit balances between the two groups. Accounts that defaulted tend to have lower credit limits, as indicated by the red distribution peaking at lower limit balance values, suggesting a potential correlation between lower credit limits and a higher risk of default. In contrast, the blue distribution, representing non-default accounts, exhibits a broader spread with a tendency towards higher limit balances, implying that accounts with higher credit limits are less likely to default. This visual comparison underscores the significance of credit limit as a factor in assessing default risk, with lower limits potentially signaling a higher propensity for financial distress.

Heatmap



The heatmap of the correlation matrix provides a comprehensive overview of the interrelationships among variables in the dataset, revealing both positive and negative correlations with varying degrees of intensity. Stronger positive correlations are represented by deeper shades of red, indicating a direct relationship where an increase in one variable is associated with an increase in another. Conversely, deeper shades of blue denote stronger negative correlations, suggesting an inverse relationship between variables. This visual tool is particularly effective for identifying potential patterns, dependencies, and contrasts within the data, serving as a foundational step for more detailed statistical analysis or predictive modeling.

Data pre processing

In the pre-processing and training data development phase, a comprehensive approach was taken to prepare a dataset for machine learning analysis. Initially, the data underwent cleaning and standardization, with columns being renamed for clarity, particularly those related to repayment statuses, bill amounts, and previous payments over several months. This step ensured a clear understanding of the dataset's features. The phase included an in-depth exploration of the dataset, identifying its size and the nature of its variables, which ranged from integers and floats to categorical types. Special attention was given to converting certain variables, like education and marital status, into categorical formats to better reflect their non-numeric nature. A significant portion of this phase was dedicated to feature engineering, where techniques like one-hot encoding were employed to transform categorical variables into a machine-readable format. This was crucial for accommodating algorithms that require numerical input. Additionally, normalization processes were applied to scale the data, addressing the needs of algorithms sensitive to variable scales and aiding in efficient optimization. The dataset was then strategically divided into training and test sets, with a focus on maintaining a representative distribution of the target variable, 'Default', in both subsets. This careful separation is fundamental to ensuring that the model can be trained effectively and evaluated accurately.

Finally, the prepared data, now split into training and testing sets, was saved for future use, marking the completion of a thorough data preparation stage. This phase set a solid foundation for the subsequent modeling efforts, emphasizing the importance of meticulous data handling and strategic preparation in enhancing model accuracy and reliability in financial outcome predictions.

Modeling

In addressing the challenge of predicting credit card default risk, our approach was to employ a variety of machine learning models. Each model was chosen for its unique strengths and ability to uncover different aspects of the underlying data patterns. The models included Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.

Model Implementations and Performance

1. Logistic Regression

Overview: Logistic Regression, despite its simplicity, is a powerful linear model for binary classification tasks. It's particularly useful for understanding the impact of individual features on the likelihood of default.

Implementation: We adjusted the class weights to counteract class imbalance, enhancing the model's sensitivity towards the minority class (default cases).

Performance: While Logistic Regression provided a solid baseline, its linear nature limited its ability to capture the complex nonlinear relationships within the data.

2. Decision Tree

Overview: Decision Trees offer an intuitive model structure that can capture nonlinear relationships through hierarchical decision rules. They are easily interpretable and can handle both numerical and categorical data.

Implementation: The model was configured to balance class weights, similar to Logistic Regression, to **address** the issue of class imbalance.

Performance: The Decision Tree model provided insights into the hierarchical importance of features but was prone to overfitting, especially with the dataset's complex feature interactions.

3. Random Forest

Overview: Random Forest builds on the concept of Decision Trees but introduces randomness in feature selection and bootstrapping of data samples to create an ensemble of trees. This randomness helps in improving model robustness and reducing overfitting.

Implementation: We employed class weights to improve model sensitivity and used hyperparameter tuning to find the optimal configuration for the ensemble.

Performance: Random Forest demonstrated significant improvement over Decision Tree by capturing more complex patterns without overfitting, making it a strong contender in our model lineup.

4. Gradient Boosting

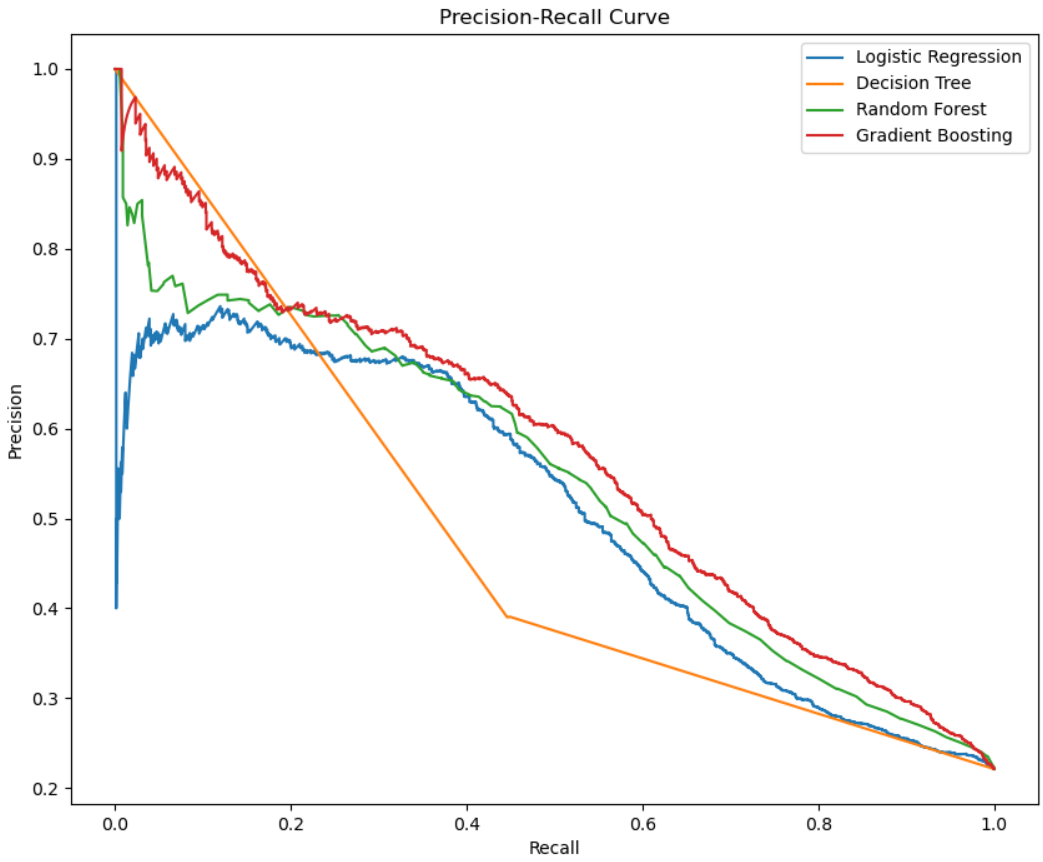
Overview: Gradient Boosting is an ensemble technique that builds trees sequentially, with each tree trying to correct the errors of its predecessor. This approach can capture very complex relationships in the data.

Implementation: Subsampling was used to prevent overfitting, and extensive hyperparameter tuning was conducted to optimize performance.

Performance: Gradient Boosting emerged as the top-performing model, showcasing superior ability to balance precision and recall, which was reflected in its high AUC-PR score. Its strength lies in its ability to leverage weak learners to build a robust predictive model.

Model selection

Figure 11: Precision-Recall Curves



Precision-Recall curves are a valuable tool for evaluating the performance of classification models, particularly in imbalanced datasets where one class is much more prevalent than the other. Here are some key insights we can derive from these curves:

Logistic Regression: This model seems to offer a reasonable balance between precision and recall, especially in the mid-range of recall values. However, it might not be the best for very high recall requirements, as precision drops off.

Decision Tree: This model's curve suggests it might not perform as well as the others, especially at higher recall levels where its precision drops significantly. It might be overfitting to the training data, a common issue with decision trees.

Random Forest: The curve suggests that this model generally performs better than the Decision Tree, likely due to its ensemble approach that helps reduce overfitting. It maintains a higher level of precision across a broad range of recall values.

Gradient Boosting: This model's curve appears to be among the best, indicating a strong performance across a wide range of recall levels. It maintains higher precision for a given recall level, which is indicative of a well-performing model.

Figure 12: AUC-PR

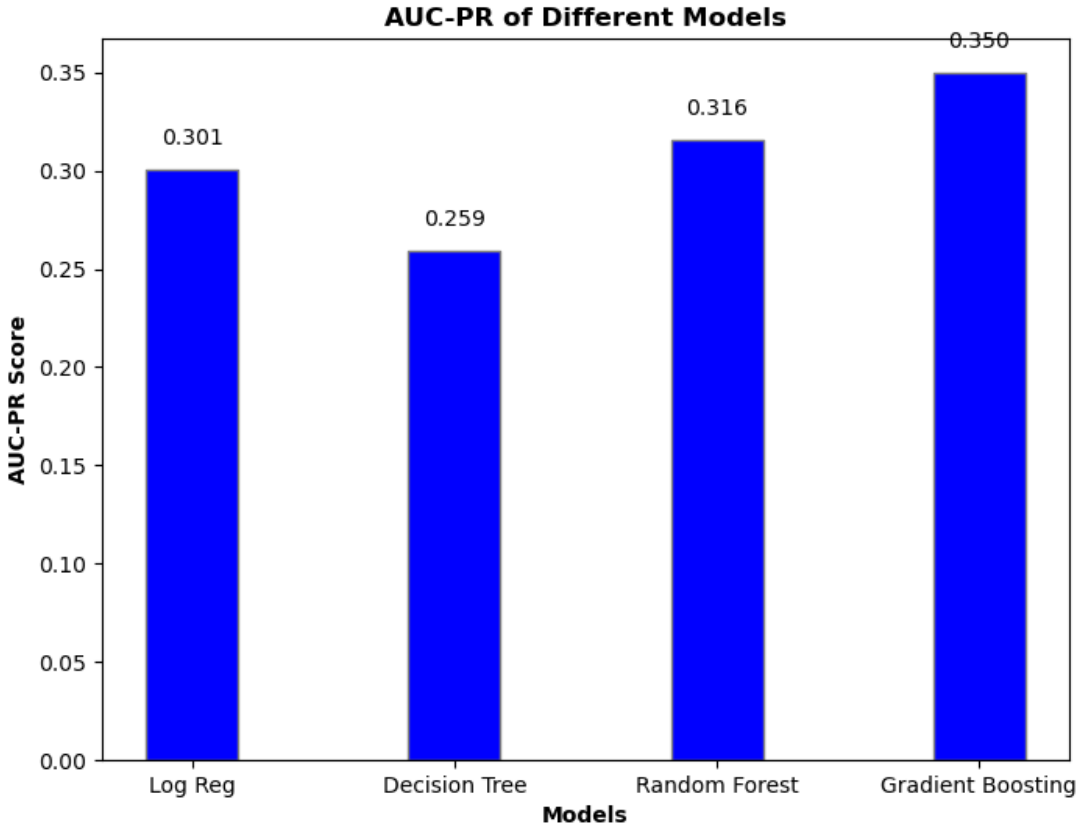
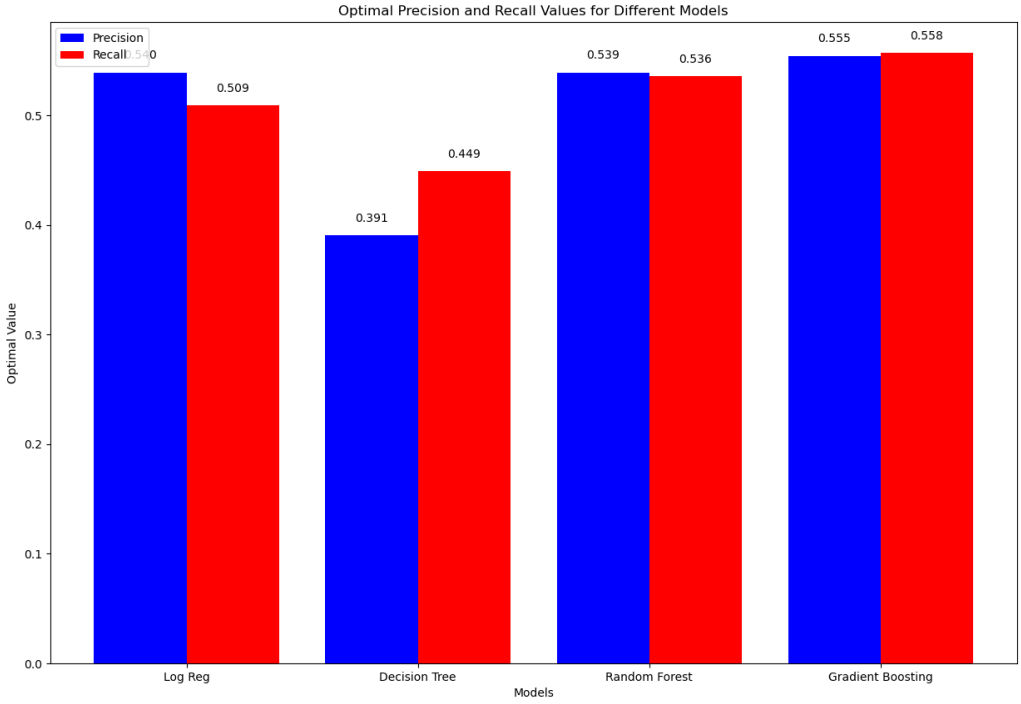


Figure 13: Optimal Precision-Recall



After conducting an extensive evaluation of various classification models on the dataset, we have identified the top-performing models based on multiple performance metrics.

- AUC-PR (Area Under the Precision-Recall Curve)

The AUC-PR scores provide insights into the overall performance of each model in terms of precision and recall. Among the models evaluated, the Gradient Boosting and Random Forest models demonstrated the highest AUC-PR scores. Specifically, the Gradient Boosting model achieved an AUC-PR score of 0.3519, while the Random Forest model achieved a score of 0.3177.

- Optimal Precision and Recall

Additionally, we examined the optimal precision and recall values for each model. These values represent the best trade-off between precision and recall, as determined by the F1 score. The Gradient Boosting model exhibited the best combination of precision and recall, with optimal precision of 0.5745 and optimal recall of 0.5463.

Taking into account both the AUC-PR scores and the optimal precision-recall values, we conclude that the Gradient Boosting and Random Forest models are the top-performing models for this classification task. These models demonstrate superior performance in terms of accurately classifying the target variable while maintaining a balance between precision and recall.

Key Takeaways

Implementing our predictive models for credit card default detection brings significant benefits to a financial institution, enhancing several aspects of its operations and strategic initiatives. These models, refined through rigorous development and validation processes, offer precise default predictions, thus significantly advancing the institution's risk management capabilities. By accurately identifying high-risk accounts, our models enable proactive risk mitigation strategies, reducing the incidence of defaults and safeguarding the institution's financial health.

The operational efficiency of the institution is notably improved with the integration of our models. The automation of risk assessments streamlines decision-making processes, minimizing the need for time-consuming manual reviews. This leads to a more judicious allocation of human resources, focusing expert analysis on complex or high-risk cases flagged by our predictive models, thus optimizing operational workflows.

From a financial perspective, the early detection capabilities of our models are instrumental in minimizing losses associated with credit defaults. By identifying potential defaults well in advance, the institution can engage in early intervention strategies, thereby protecting its revenue streams. Additionally, the insights generated by our models can be leveraged to identify opportunities for safe revenue growth, such as extending credit offers to identified low-risk segments, thereby enhancing profitability. The customer experience is also significantly enhanced by the application of our predictive models. The nuanced understanding of risk profiles facilitated by our models allows for the development of customized credit products and services, tailored to meet the diverse needs of the customer base. Furthermore, early identification of customers at risk of default enables the institution to provide targeted support and solutions, potentially averting defaults and fostering stronger customer loyalty. On a strategic level, our models serve as a powerful tool for data-driven decision-making. The insights provided by the models can inform key strategic initiatives, such as market expansion or the development of new credit products, aligning closely with the institution's long-term goals. The competitive edge afforded by our advanced predictive capabilities enhances the institution's position in the market, driving innovation in product offerings and risk management strategies.

In summary, the deployment of our predictive models for credit card default detection offers a comprehensive suite of benefits, encompassing improved risk management, operational efficiencies, financial performance, customer satisfaction, strategic decision support, and regulatory compliance. These models greatly improve how the institution manages credit risk, leading to steady growth and a stronger position in the market.

Future work

Looking ahead, the continuous refinement and application of our predictive models in credit risk management hold the promise of even greater benefits. As we gather more data and refine our algorithms, we can expect our models to become more accurate and insightful, allowing for even more precise risk assessments and proactive interventions.

Future advancements in technology and analytics may also introduce new opportunities to enhance our models, such as integrating artificial intelligence and machine learning techniques that can learn and adapt in real-time to emerging trends and patterns. This could lead to more personalized and dynamic risk management strategies, further reducing defaults and enhancing customer relationships.

Additionally, the integration of our predictive models with other financial products and services could open up new avenues for innovation, offering customers a more holistic and responsive financial management experience. This could not only improve customer satisfaction and loyalty but also attract new customers looking for a more data-driven and personalized approach to financial services.

In the broader financial landscape, as regulatory and market conditions evolve, our models can provide a solid foundation for agility and compliance, ensuring that the institution remains at the forefront of responsible and innovative credit management practices. This forward-looking approach will be key to sustaining growth and maintaining a competitive edge in an increasingly complex and competitive market.

table of contents

List of figures